W1200

DISK ARRAY CONTROLLER

Patent number: JP10333836
Publication date: 1998-12-18
Inventor: FUJIMOTO KAZUHISA; TANAKA ATSUSHI; ODAWARA HIROAKI
Applicant: HITACHI LTD
Classification:
- international: G06F3/06; G06F3/06; G11B20/10; G11B20/18; G11B20/18
- european:
Application number: JP19970139656 19970529
Priority number(s):

View INPADOC patent family

Abstract of JP10333836

PROBLEM TO BE SOLVED: To reduce the communication load on a shared memory part, and to improve the total throughput of a disk array controller, by collecting plural transmission data having the same sending destination in a single composite packet. SOLUTION: The control information sent from a microprocessor 100 or the data sent from a data transmission/reception control part 110, which performs the transmission/reception of data to a host computer are sent to a communication controller 140 and then stored in the buffers 130 and 131 prepared for every sending destination. When plural transmission data having the same sending destination are stored, a packet generation part 120 collects these data in a composite packet and sends it to a desired shared memory part. Then, the composite packet sent from the shared memory part in a composite packet are decomposed into plural transmission data at a packet decomposition part 125 and stored in a buffer 135. These decomposed data are sent to a microprocessor 106 or the control part 110 via the controller 140.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－333836

(43)公開日　平成10年(1998)12月18日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	3/06	３０１	Ｇ０６Ｆ	3/06	３０１Ｒ
		５４０			５４０
Ｇ１１Ｂ	20/10		Ｇ１１Ｂ	20/10	Ｄ
	20/18	５７０		20/18	５７０Ｚ
		５７２			５７２Ｆ

審査請求　未請求　請求項の数５　ＯＬ　（全７頁）

(21)出願番号　特願平9－139656

(22)出願日　平成９年(1997)５月29日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　藤本　和久
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　田中　淳
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

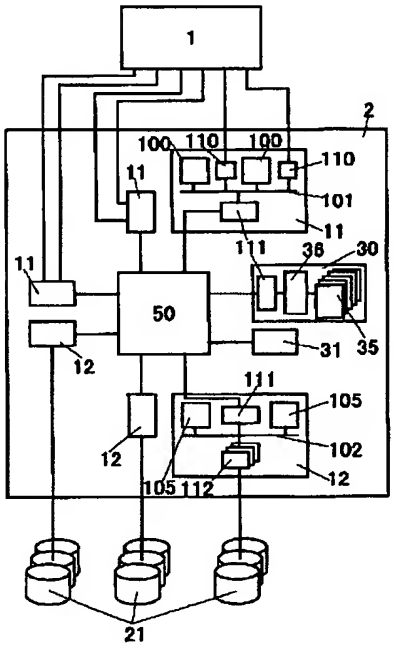
(72)発明者　小田原　宏明
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(74)代理人　弁理士　小川　勝男

(54)【発明の名称】　ディスクアレイ制御装置

(57)【要約】
【課題】ディスクアレイ装置において、高いＩ／Ｏスループット性能を実現する。
【解決手段】ホストコンピュータ１とのインターフェース部１１と、複数の磁気ディスク装置２１とのインターフェース部１２と、共有メモリ部３０，３１内の、スイッチとの通信制御部１１１において、バッファ内に格納された送出先が同一の複数の送信データがあるときは、それらを１つの複合パケットにまとめて送出する。

図1



【特許請求の範囲】
【請求項１】少なくとも、ホストコンピュータとの１つのインターフェース部と、複数の磁気ディスク装置との１つのインターフェース部と、データ及び制御情報を格納する物理的に独立した複数の共有メモリ部から成り、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記複数の共有メモリ部が、スイッチを用いた相互結合網によって結合されたディスクアレイ制御装置であって、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部において、バッファ内に格納された送出先が同一の複数の送信データを１つの複合パケットにまとめて送出することを特徴とするディスクアレイ制御装置。
【請求項２】請求項１記載のディスクアレイ制御装置であって、前記共有メモリ部内のメモリ制御部において、該共有メモリ部に送出された前記複数の送信データから成る複合パケット内の複数の送信データを並列に処理することを特徴とするディスクアレイ制御装置。
【請求項３】請求項１または２において、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部内のバッファが、それぞれの送信データの送出先ごとに、物理的あるいは論理的に分割されていることを特徴とするディスクアレイ制御装置。
【請求項４】請求項１から３において、前記ホストコンピュータとのインターフェース部、及び前記複数の磁気ディスク装置とのインタフェース部が、該インターフェース部内の処理を分散して行う複数のマイクロプロセッサから成ることを特徴とするディスクアレイ制御装置。
【請求項５】請求項１から４において、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部内のバッファからの送信データ送出処理の際、該バッファ内に、送出先が同一の送信データが少なくとも２つ以上格納されている場合にのみ、該送信データ群を１つの複合パケットにまとめて送出することを特徴とするディスクアレイ制御装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、データを分割して複数の磁気ディスク装置に格納するディスクアレイ装置の制御装置に関する。
【０００２】
【従来の技術】コンピュータの主記憶のＩ／Ｏ性能に比べて、２次記憶装置として用いられる磁気ディスク装置を使ったサブシステムのＩ／Ｏ性能は３～４桁程度小さく、従来からこの差を縮めること、すなわちサブシステムのＩ／Ｏ性能を向上する努力が各所でなされている。サブシステムのＩ／Ｏ性能を向上させるための１つの方法として、複数の磁気ディスク装置でサブシステムを構成し、データを分割して複数の磁気ディスク装置に格納する手段、いわゆるディスクアレイと呼ばれるシステムが知られている。
【０００３】このディスクアレイでは、複数の磁気ディスク装置を並列に動作させてデータの入出力を行うため、Ｉ／Ｏ性能が向上する。しかし、複数の磁気ディスク装置への書き込み及び読み出し動作を並列に処理するため、制御が複雑で、それに要するオーバーヘッドが大きいという問題がある。
【０００４】図２に主にメインフレーム向けの大型ディスクアレイ装置の制御装置３を示す。この制御装置３は一般的に、ホストコンピュータ１とのインターフェース部１３，複数の磁気ディスク装置２１とのインターフェース部１４，データ及び制御情報を一時格納する共有メモリ部３２が、共有バス６０を介して繋がる構成をとっている。
【０００５】ディスクアレイ装置のＩ／Ｏスループット性能の伸びは大きく、それに対応するため、上記インターフェース部等の処理性能を向上させる必要がある。これら処理性能の向上に伴って、共有バス６０の利用率が飽和状態となり、それが原因でスループット性能が制限されている。そこで、共有バス６０のスループットを上げるための努力がなされているが、装置の構成上、バス幅，駆動周波数等を改善することは難しく、スループットの向上にも限界がある。
【０００６】そのため、共有バス６０に代わって、スイッチを用いた相互結合網を介して上記インターフェース部等を繋ぐことが考えられている。この方法では、相互接続された個々のパスのスループットは、共有バスの数分の１であるが、相互接続された２点間には複数のパスが存在するため、負荷が分散され、スループットの向上が可能となる。
【０００７】
【発明が解決しようとする課題】スイッチを用いた相互結合網では、多対多の通信が一般的であるが、ディスクアレイの制御装置では、図２に示す複数のインターフェース部１３，１４と共有メモリ部３２間の通信が大部分を占める。したがって、負荷分散のため、共有メモリ部を物理的に複数に分割することが考えられている。しかしながら、共有メモリ部の分割数にも限界があり、均等な負荷分散を行うことは難しい。したがって、ディスクアレイ装置においては、スイッチを用いた相互結合網が持つ本来のスループット性能を引き出すことは難しく、スループット向上にも限界がある。
【０００８】本発明の目的は上述の課題を解消し、Ｉ／Ｏスループット性能の高いディスクアレイ装置を提供す

ることにある。
【０００９】
【課題を解決するための手段】上記目的は、少なくとも、ホストコンピュータとの１つのインターフェース部と、複数の磁気ディスク装置との１つのインターフェース部と、データ及び制御情報を格納する物理的に独立した複数の共有メモリ部から成り、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記複数の共有メモリ部が、スイッチを用いた相互結合網によって結合されたディスクアレイ制御装置であって、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部において、バッファ内に格納された送出先が同一の複数の送信データを１つの複合パケットにまとめて送出するディスクアレイ制御装置によって達成される。
【００１０】すなわち、送出先が同一の複数の送信データを１つの複合パケットにまとめることにより、複数回の通信オーバーヘッドが１回に減るため、１つのパケット長は長くなるが全体の通信量は減る。したがって、共有メモリ部への通信負荷が減り、制御装置全体のスループットの向上が可能となる。
【００１１】また、共有メモリ部内のメモリ制御部で、その共有メモリ部に送出された複数の送信データから成る複合パケット内の複数の送信データを並列に処理することによって、共有メモリ部のスループットの向上が可能となる。これにより、制御装置全体のスループットがさらに向上する。
【００１２】また、ホストコンピュータとのインターフェース部と、複数の磁気ディスク装置とのインターフェース部と、共有メモリ部内の、スイッチとのインターフェース部内のバッファを、送信データそれぞれの送出先ごとに物理的あるいは論理的に分割することにより、送出先が同一の送信データを１つの複合パケットにまとめる処理が簡単になる。これにより、スイッチとのインターフェース部での処理オーバーヘッドの削減が可能となる。
【００１３】ホストコンピュータとのインターフェース部、及び複数の磁気ディスク装置とのインターフェース部が、そのインターフェース部内の処理を分散して行う複数のマイクロプロセッサから成っている場合、複数のマイクロプロセッサが同一の共有メモリ部にほぼ同時にアクセスすることが発生するため、スイッチとのインターフェース部のバッファ内に、送出先が同一の送信データが複数存在する頻度が高くなる。したがって、この場合、上記手段は特に有効となる。
【００１４】スイッチとのインターフェース部内のバッファからの、送信データ送出処理の際、該バッファ内に送出先が同一の送信データがない場合に、送出先が同一の送信データ発生を待つと、待ち時間が長くなった場合に、制御装置全体の応答性能が悪くなる可能性がある。そのため、送出先が同一の送信データが少なくとも２つ以上格納されている場合にのみ、該送信データ群を１つの複合パケットにまとめて送出することが望ましい。
【００１５】
【発明の実施の形態】
（実施例１）図１に、本発明の制御装置の一実施例を示す。ディスクアレイ制御装置２内の、ホストコンピュータ１とのインターフェース部１１と、複数の磁気ディスク装置２１とのインターフェース部１２と、２つの共有メモリ部３０，３１は、スイッチを用いた相互結合網５０を介して結合されている。
【００１６】スイッチを用いた相互結合網５０は、複数のポートを有する少なくとも１つのスイッチから成る。各スイッチのポートには、他のスイッチ、あるいはディスクアレイ制御装置内の各インターフェース部及び共有メモリ部が繋がる。本実施例では、図７に示すように入出力にそれぞれ５つのポートを有するスイッチ４０を４つ、相互に結合した（完全結合網と呼ぶトポロジー）相互結合網５０を用いる。図では、入出力ポートをまとめて１つの線として示している。他のトポロジーの相互結合網を用いても本実施例を実施する上で問題はない。
【００１７】ホストコンピュータ１とのインターフェース部１１内では、２つの制御用マイクロプロセッサ１００，ホストコンピュータとの２つのデータ送受信制御部110、及びインターフェース部－共有メモリ部間の通信制御部１１１が共有バス101を介して接続されている。
【００１８】複数の磁気ディスク装置２１とのインターフェース部１２内では、２つの制御用マイクロプロセッサ１０５，複数の磁気ディスク装置２１とのデータ送受信制御部１１２、及びインターフェース部－共有メモリ部間の通信制御部１１１が共有バス１０２を介して接続されている。
【００１９】共有メモリ部３０，３１内では、インターフェース部－共有メモリ部間の通信制御部１１１を介して、メモリ制御部３６，メモリモジュール３５が接続されている。
【００２０】ホストコンピュータ１からディスクアレイ制御装置２へのデータの読み出し要求は、ホストコンピュータとのデータの送受信制御部１１０を通して、マイクロプロセッサ１００に伝える。マイクロプロセッサ１００は、共有メモリ３０，３１内に要求データが有るかどうかを確認するため、インターフェース部－共有メモリ部間の通信制御部１１１を通して通信を行う。要求データが存在する場合、マイクロプロセッサ１００は、共有メモリからの読み出し処理を実行する。要求データが存在しない場合、マイクロプロセッサ１００は、磁気ディスク装置２１から共有メモリ部３０あるいは３１への要求データの転送命令を、共有メモリ部３０あるいは３

１，目的ポートNo.（送出先のＩＤ番号）２６２，ルーティング情報（途中経由するスイッチの番号）２６３から成る。これらの内容を経由スイッチ内で参照し、パケットを目的のポートまで届ける。この方法以外に、ルーティング情報２６３無しに、スイッチ内で次に経由するスイッチを動的に割り当てることによって、パケットを目的ポートへ届ける方法もある。この方法によっても、本発明を実施する上で問題はない。

【００３１】図５は、複数の磁気ディスク装置２１とのインターフェース部１２内の、スイッチとの通信制御部１１１の構成を示している。インターフェース部１２内では、２つの制御用マイクロプロセッサ１０５，複数の磁気ディスク装置２１とのデータ送受信制御部１１２、及びインターフェース部ー共有メモリ部間の通信制御部１１１が共有バス１０２を介して接続されている。このインターフェース部１２においても、図３，図４において説明した方法と同様の制御を、スイッチとの通信制御部１１１において行う。

【００３２】図６は、共有メモリ部３０，３１内の、スイッチとの通信制御部１１１の構成を示している。共有メモリ部３０，３１内では、インターフェース部ー共有メモリ部間の通信制御部１１１を介して、メモリ制御部３６，メモリモジュール３５が接続されている。この共有メモリ部３０，３１においても、図３，図４において説明した方法と同様の制御を、スイッチとの通信制御部１１１において行う。

【００３３】以上述べたように、送出先が同一の複数の送信データを１つの複合パケットにまとめることにより、複数回の通信オーバーヘッドが１回に減るため、１つのパケット長は長くなるが全体の通信量は減る。したがって、共有メモリ部への通信負荷の削減が可能となる。

【００３４】ここでは、ホストコンピュータ１とのインターフェース部１１、及び複数の磁気ディスク装置２１とのインタフェース部１２が、そのインターフェース部内の処理を分散して行う２つのマイクロプロセッサ１００あるいは１０５から成っている場合について説明したが、インターフェース部１１あるいは１２が、１つのマイクロプロセッサから成っている場合も、本発明を実施する上で問題はない。しかしながら、インターフェース部１１あるいは１２が複数のマイクロプロセッサから成っている場合、それらが同一の共有メモリ部にほぼ同時にアクセスすることが発生するため、スイッチとのインターフェース部のバッファ内に、送出先が同一の送信データが複数存在する頻度が高くなる。したがって、この場合、本発明は特に有効である。

【００３５】図８は、スイッチによる相互結合網を用いた従来のディスクアレイ制御装置と本発明のディスクアレイ装置のスループット性能を計算によって求め、比較した結果を示している。ホストからの負荷の条件は、ベンチマークプログラムＰＡＩ－ＩＯ－Driverのzero－Locality での負荷条件（Read：Write比が１：１で、キャッシュヒット率が０％）とした。ディスクアレイ制御装置は、ホストコンピュータとのインターフェース部１１のパッケージ８枚と複数の磁気ディスク２１とのインターフェース部１２のパッケージが１６枚，共有メモリ部のパッケージが８枚から構成されているとした。

【００３６】図では、縦軸に共有メモリへのアクセス時間を、限界のアクセス時間を１とした相対値で示している。また、横軸にスループットを、従来の限界値を１とした相対値で示している。図からわかるように、従来に比べて本発明のディスクアレイ制御装置では、スループットが約２６％向上する。

【００３７】ディスクアレイ制御装置２では、各インターフェース部１１，１２と共有メモリ部３０，３１との間の通信が中心となってホストコンピュータ１からのＩ／Ｏ要求の処理が行われる。したがって、ホストコンピュータからのＩ／Ｏ負荷が増加するにつれて、上記の通信量が増大する。その結果、スイッチを用いた相互結合網の利用率が飽和状態となり、スループット性能が制限される。しかしながら、本実施例によれば、共有メモリ部への通信負荷が削減でき、スイッチを用いた相互結合網全体のスループットが上がるため、ディスクアレイ制御装置２全体のスループット向上が可能となる。

【００３８】（実施例２）本発明の他の一実施例を示す。実施例１で述べたディスクアレイ制御装置２において、共有メモリ部３０，３１内のメモリ制御部３６で、その共有メモリ部に送出された複数の送信データから成る複合パケット内の複数の送信データを並列に処理する機能を持たせる。すなわち、図９に示すように、複数のメモリモジュール３５へのパスを複数設け、それぞれのパス毎に制御コントローラ３７を設けることによって、複数の送信データを並列に処理する。

【００３９】本実施例によれば、共有メモリ部のスループットが上がるため、ディスクアレイ制御装置全体のスループットをさらに向上することが可能となる。

【００４０】（実施例３）本発明の他の一実施例を示す。実施例１または２で述べたディスクアレイ制御装置２において、スイッチとの通信制御部１１１のバッファ内からの送信データ送出処理の際、そのバッファ内に送出先が同一の送信データがない場合に送出先が同一の送信データ発生を待つと、待ち時間が長くなった場合に制御装置全体の応答時間性能が悪くなる場合があると考えられる。そこで、スイッチとの通信制御部１１１で、送出先が同一の送信データが少なくとも２つ以上格納されている場合にのみ、該送信データ群を１つの複合パケットにまとめて送出する。送信データが１つしかない場合には、それのみをパケットとして送出するという制御を行う。

【００４１】本実施例によれば、複合パケットによる共

有メモリ部との通信を行った場合のディスクアレイ制御装置全体の応答時間性能の低下を防ぐことが可能となる。

【００４２】

【発明の効果】本発明によれば、スイッチを用いた相互結合網内において共有メモリ部への負荷が減るため、そのスループットが向上する。それによって、ディスクアレイ制御装置の各インターフェース部と共有メモリ部間の通信の飽和が解消され、装置全体のスループット性能が向上する。

【図面の簡単な説明】

【図１】本発明の一実施例のディスクアレイ制御装置の構成を示すブロック図。

【図２】従来のディスクアレイ制御装置の構成を示すブロック図。

【図３】図１のホストコンピュータとスイッチとのインターフェース部の構成を示すブロック図。

【図４】本発明の一実施例のパケットのフォーマットを示す説明図。

【図５】図１の磁気ディスク装置とスイッチとのインターフェース部の構成を示すブロック図。

【図６】図１の共有メモリとスイッチとのインターフェース部の構成を示すブロック図。

【図７】図１のスイッチを用いた相互結合網の構成を示すブロック図。

【図８】従来と本発明によるディスクアレイ制御装置のスループット性能の予測結果を比較した図。

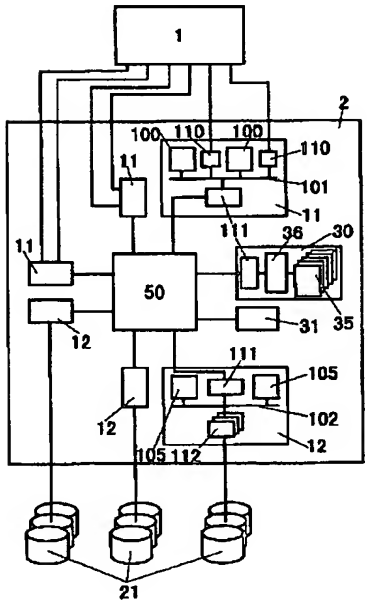
【図９】本発明の他の実施例における共有メモリ部の構成を示すブロック図。

【符号の説明】

１…ホストコンピュータ、２…ディスクアレイ制御装置、１１…ホストコンピュータとのインターフェース部、１２…磁気ディスク装置とのインターフェース部、２１…磁気ディスク装置、３０…共有メモリ部、３５…メモリモジュール、３６…メモリ制御部、５０…スイッチを用いた相互結合網、１００…マイクロプロセッサ、１０１…共有バス、１０２…共有バス、１０５…マイクロプロセッサ、１１０…ホストコンピュータとのデータの送受信制御部、１１１…通信制御部、１１２…磁気ディスク装置とのデータ送受信制御部、１２０…パケット生成部、１２５…パケット分解部、１３０，１３１…送信先別バッファ、１３５…バッファ、１４０…通信制御コントローラ。

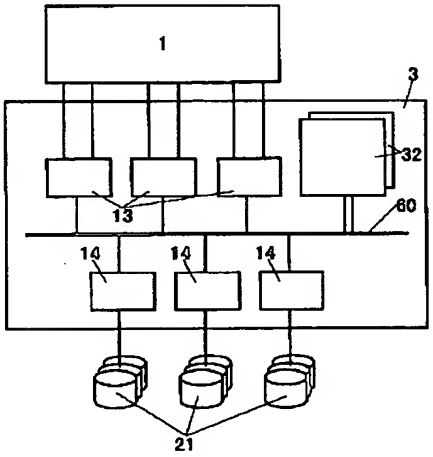
【図１】

図1



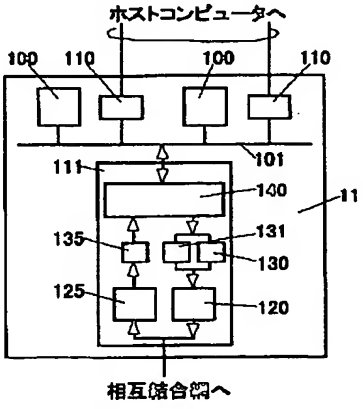
【図２】

図2



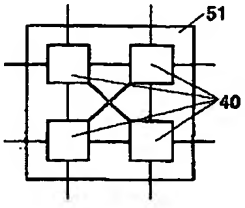
【図３】

図3



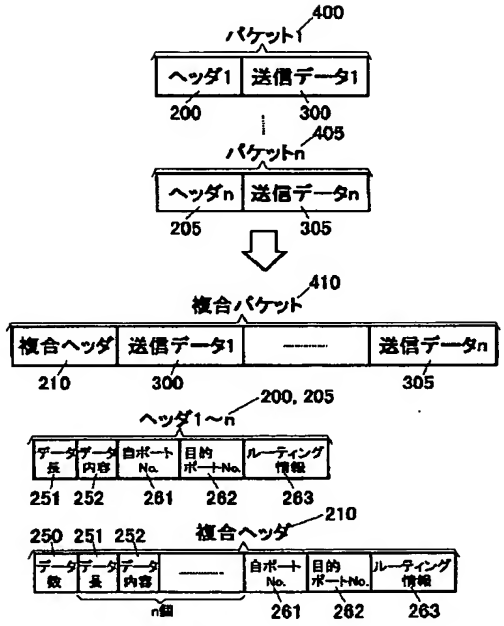
【図７】

図7



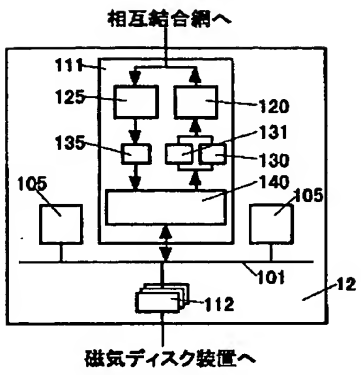
【図４】

図4



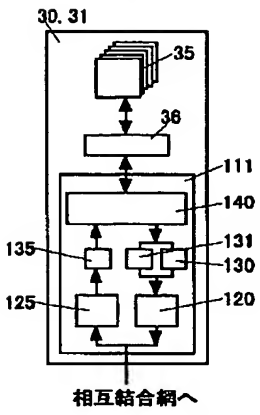
【図５】

図5



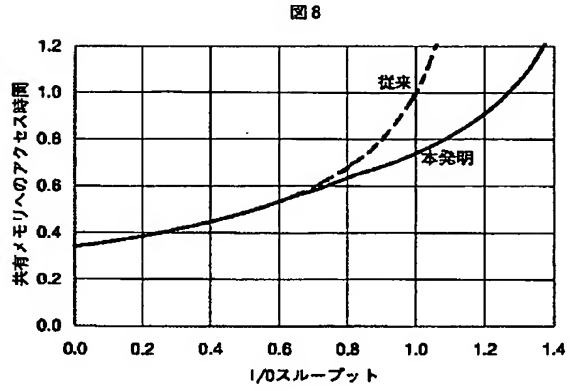
【図６】

図6



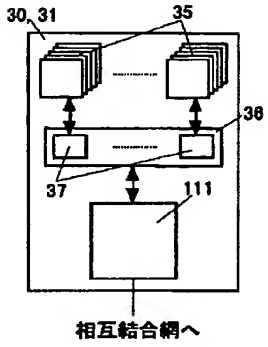
【図８】

図8



【図９】

図9



US 1070048502P1

Creation date: 02-11-2004
Indexing Officer: KSAM - KONA SAM
Team: OIPEScanning
Dossier: 10700485

Legal Date: 01-09-2004

No.	Doccode	Number of pages
1	IDS	9
2	NPL	4
3	NPL	4
4	NPL	7
5	NPL	7

Total number of pages: 31

Remarks:

Order of re-scan issued on .....